On modifications of some distance based two-sample tests for high dimensional data

Soham Sarkar and Anil K. Ghosh

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203, B. T. Road, Kolkata 700108, India.

Email: sohamsarkar1991@gmail.com, akghosh@isical.ac.in

Abstract

Over the last two decades, several two-sample tests based on pairwise Euclidean distances have been proposed in the literature. Most of these tests can be used even when the dimension of the data is much larger than the sample size. For instance, the rigid motion invariant tests proposed by Baringhaus and Franz (2010) and those proposed by Biswas and Ghosh (2014) can be conveniently used in high dimension, low sample size (HDLSS) situations. In this article, we show that these tests can yield poor results in the HDLSS set up even when the two distributions differ widely in their scatters and shapes. To overcome this limitation, we propose some simple modifications of these tests, where the Euclidean distance is replaced by other appropriate distance functions. Similar modification is carried out for graph based multivariate run tests as well. We prove the high dimensional consistency of these modified tests under appropriate regularity conditions and evaluate their empirical performance using several simulated and real data sets.

Keywords: Cramér test, HDLSS asymptotics, multivariate run test, permutation test.

1 Introduction

In a two-sample problem, we test for the equality of two d-dimensional distributions F and G based on two independent samples $\mathbf{x}_1, \ldots, \mathbf{x}_m$ from F and $\mathbf{y}_1, \ldots, \mathbf{y}_n$ from G. This problem is well investigated in the literature, and several tests are available for it. In the parametric regime, we have the Hotelling's T^2 test (see e.g., Anderson, 2003), which assumes F and G to be Gaussian having the same dispersion structure and tests for the equality of their locations. Nonparametric tests for the two-sample location problem include Puri and Sen (1971); Randles and Peters (1990); Hettmansperger and Oja (1994); Choi and Marden (1997) and Hettmansperger et al. (1998). But many of these tests become computationally prohibitive even for moderately high dimensional data, and none of them can be used when the dimension exceeds the sample size. Two-sample location tests that can be used for high dimension, low sample size (HDLSS) data include Bai and Saranadasa (1996); Chen and Qin (2010); Srivastava et al. (2013); Park and Ayyala (2013); Ghosh and Biswas (2016); Wei et al. (2016). Several tests are available for the general two-sample problem as well, where one tests the null hypothesis $\mathcal{H}_0 : F = G$ against the alternative $\mathcal{H}_A : F \neq G$ (see, e.g., Friedman and Rafsky, 1979; Schilling, 1986; Henze, 1988; Ferger, 2000; Hall and Tajvidi, 2002; Rosenbaum, 2005; Liu and Modarres, 2011; Gretton et al., 2012; Biswas and Ghosh, 2014; Biswas et al., 2014; Mondal et al., 2015).

Many of these above mentioned tests use test statistics based on pairwise Euclidean distances. It is known that if $\mathbf{X}_1, \mathbf{X}_2$ are two independent copies of $\mathbf{X} \sim F$ and $\mathbf{Y}_1, \mathbf{Y}_2$ are two independent copies of $\mathbf{Y} \sim G$, then $\|\mathbf{X}_1 - \mathbf{X}_2\|$, $\|\mathbf{X}_1 - \mathbf{Y}_1\|$ and $\|\mathbf{Y}_1 - \mathbf{Y}_2\|$ have the same distribution if and only if F and G are identical (Maa et al., 1996). So, pairwise Euclidean distances contain useful information about the difference between two multivariate distributions. Under the usual moment conditions, we also have $2E\|\mathbf{X}_1 - \mathbf{Y}_1\| - E\|\mathbf{X}_1 - \mathbf{X}_2\| - E\|\mathbf{Y}_1 - \mathbf{Y}_2\| \geq 0$, where the equality holds if and only if $F = G$. Motivated by this result, Baringhaus and Franz (2004) constructed the Cramér test, which rejects $\mathcal{H}_0$ for large values of the test statistic

$$T_C = \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\|\mathbf{x}_i - \mathbf{y}_j\| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|\mathbf{x}_i - \mathbf{x}_j\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{y}_i - \mathbf{y}_j\|.$$

This test has the large sample consistency against general alternatives. Under usual moment conditions, its power converges to unity as the sample size increases. From the expression of T_C, it is also clear that this test can be conveniently used for high dimensional data even when the dimension is much larger than the sample size. But, Biswas and Ghosh (2014) showed that it often fails to perform well it the HDLSS set up, particularly when the underlying distributions differ mainly in their scales. To take care of this problem, they proposed to use

$$T_{BG} = \Big|\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\|\mathbf{x}_i - \mathbf{y}_j\| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|\mathbf{x}_i - \mathbf{x}_j\|\Big|^2 + \Big|\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\|\mathbf{x}_i - \mathbf{y}_j\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{y}_i - \mathbf{y}_j\|\Big|^2$$

as the test statistic. Baringhaus and Franz (2010) also considered a class of rigid motion invariant test statistics, where $\|\cdot\|$ in T_C was replaced by $\phi(\|\cdot\|^2)$, for $\phi : [0,\infty) \to [0,\infty)$ being a suitably chosen function. In particular, they chose $\phi(t) = \sqrt{t}$ (which leads to the Cramér test), $\log(1+t), t^2/(1+t^2)$ and $\exp(t) - 1$. A similar class of rigid motion invariant tests can be constructed if $\|\cdot\|$ in T_{BG} is replaced by $\phi(\|\cdot\|^2)$. However, one should note that $\|\cdot\|^2$ usually diverges to infinity at the rate of $\mathbf{O}(d)$ (see, e.g., Hall et al., 2005). Therefore, to use these tests meaningfully for high dimensional data, instead of $\phi(\|\cdot\|^2)$, we use $\phi(\|\cdot\|^2/d)$ and consider test statistics of the form

$$T_C^{\phi} = 2\widehat{\mu}_{\phi}(F,G) - \widehat{\mu}_{\phi}(F,F) - \widehat{\mu}_{\phi}(G,G) \text{ and } T_{BG}^{\phi} = \{\widehat{\mu}_{\phi}(F,G) - \widehat{\mu}_{\phi}(F,F)\}^2 + \{\widehat{\mu}_{\phi}(F,G) - \widehat{\mu}_{\phi}(G,G)\}^2,$$

where $\widehat{\mu}_{\phi}(F,F) = m^{-2}\sum_{i=1}^{m}\sum_{j=1}^{m}\phi(\|\mathbf{x}_i - \mathbf{x}_j\|^2/d)$, $\widehat{\mu}_{\phi}(G,G) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi(\|\mathbf{y}_i - \mathbf{y}_j\|^2/d)$ and $\widehat{\mu}_{\phi}(F,G) = (mn)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi(\|\mathbf{x}_i - \mathbf{y}_j\|^2/d)$. Note that for $\phi(t) = \sqrt{t}$, we have $T_C^{\phi} = T_C/\sqrt{d}$ and $T_{BF}^{\phi} = T_{BF}/d$. Henceforth, the first class of tests based on T_C^{ϕ} will be referred to as Cramér tests, while the class of tests based on T_{BG}^{ϕ} will be referred to as BG tests. Biswas and Ghosh (2014)

showed that Cramér tests often fail to perform well in high dimensional data, especially when the two distributions differ only in their scales. In such cases, BG tests are preferred. But these tests also have their limitations in the HDLSS set up. To demonstrate this, we consider the following two examples.

Example 1 : Each of the two distributions F and G are Gaussian with the mean vector $\mathbf{0}_d = (0, \ldots, 0)^\top$ and diagonal scatter matrices. The first $d/2$ diagonal elements of the scatter matrix of F (respectively, G) are 1 (respectively, 1.5) and the rest are 1.5 (respectively, 1).

Example 2 : Both F and G have independent and identically distributed measurement variables. For F, they are distributed as $\mathcal{N}(0, 5)$, while for G, they have the $t_5(0, 3)$ distribution, where $t_\nu(\mu, \sigma^2)$ denotes a t-distribution with ν degrees of freedom, location μ and scale σ.

For each of these examples, we generated 20 observations from each distribution, and tests based on T_C^ϕ and T_{BG}^ϕ were used with $\phi(t) = \sqrt{t}, \log(1+t)$ and $\exp(t)-1$. We will refer to them as $T_C^{sqrt}, T_C^{\log}, T_C^{\exp}$ and $T_{BG}^{sqrt}, T_{BG}^{\log}, T_{BG}^{\exp}$, respectively. Each experiment was repeated 500 times, and the power of a test was computed as the proportion of times it rejected $\mathcal{H}_0$.

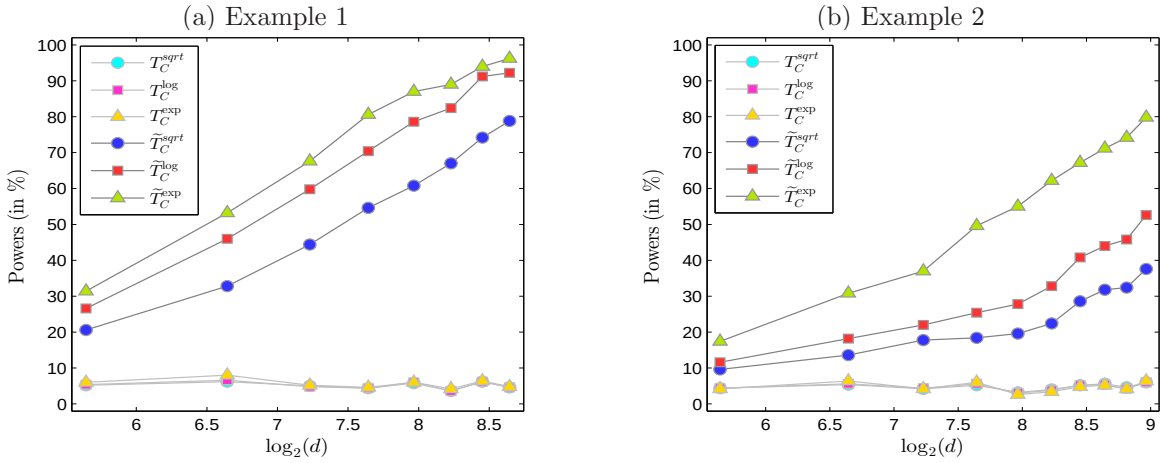

Figure 1: Powers of Cramér tests and their modified versions based on $\varphi_{h,\psi}$.

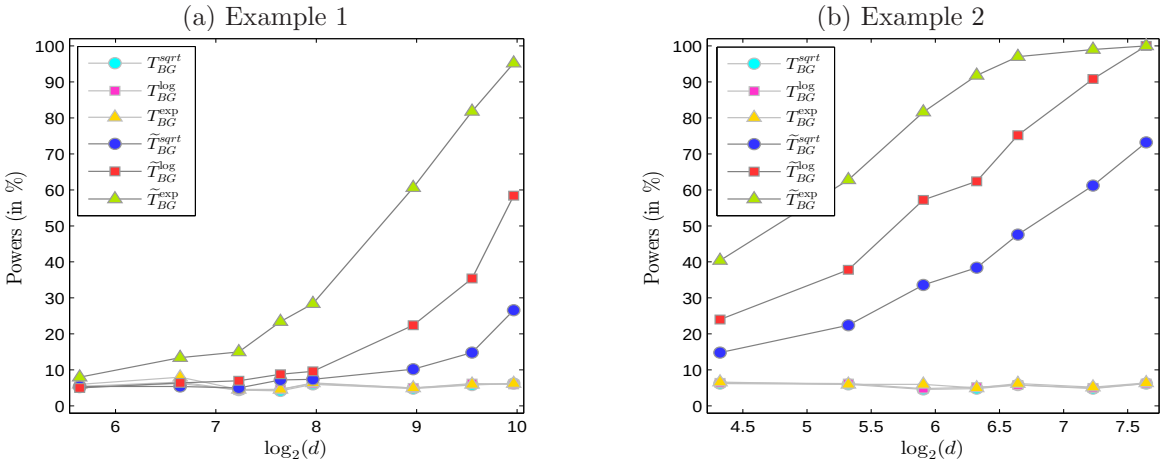

Figure 2: Powers of BG tests and their modified versions based on $\varphi_{h,\psi}$.

Figures 1 and 2 show the observed powers of these tests for different values of d. Note that in these examples, each measurement variable contains some signal against $\mathcal{H}_0$. So, one can expect the power of a test to increase as d increases. But that was not the case for the tests based on T_C^{ϕ} and T_{BG}^{ϕ}. They had poor performance in these examples. But when modified versions of these tests based on other appropriate distance functions were used, they had much higher powers (see Figures 1 and 2). Descriptions of these modified tests are given in the next section, but before that we investigate the reasons behind the failure of the tests based on T_C^{ϕ} and T_{BG}^{ϕ}.

2 Modifications of Cramér and BG tests

Let $\mathbf{X}_1, \mathbf{X}_2$ be two independent copies of $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})^\top \sim F$ and $\mathbf{Y}_1, \mathbf{Y}_2$ be two independent copies of $\mathbf{Y} = (Y^{(1)}, \ldots, Y^{(d)})^\top \sim G$. In Examples 1 and 2, all measurement variables $X^{(1)}, \ldots, X^{(d)}$ and $Y^{(1)}, \ldots, Y^{(d)}$ are independently distributed with bounded variance. Therefore, as d increases, $Var(d^{-1}\|\mathbf{X}_1 - \mathbf{X}_2\|^2) = Var\{d^{-1}\sum_{q=1}^{d}(X_1^{(q)} - X_2^{(q)})^2\}$ converges to 0 and hence $d^{-1}\big|\|\mathbf{X}_1 - \mathbf{X}_2\|^2 - E\|\mathbf{X}_1 - \mathbf{X}_2\|^2\big| \xrightarrow{P} 0$. So, using the continuous mapping theorem, one gets $\big|\phi(d^{-1}\|\mathbf{X}_1 - \mathbf{X}_2\|^2) - \phi(d^{-1}E\|\mathbf{X}_1 - \mathbf{X}_2\|^2)\big| \xrightarrow{P} 0$ as d tends to infinity. Similarly, $\big|\phi(d^{-1}\|\mathbf{Y}_1 - \mathbf{Y}_2\|^2) - \phi(d^{-1}E\|\mathbf{Y}_1 - \mathbf{Y}_2\|^2)\big| \xrightarrow{P} 0$ and $\big|\phi(d^{-1}\|\mathbf{X}_1 - \mathbf{Y}_1\|^2) - \phi(d^{-1}E\|\mathbf{X}_1 - \mathbf{Y}_1\|^2)\big| \xrightarrow{P} 0$ as d increases. This type of convergence holds even when the measurement variables are not independent, but in that case, we need some additional conditions on the underlying distributions. Some sufficient conditions in this context are given below.

(A1) Fourth moments of the measurement variables are uniformly bounded.

(A2) For $\mathbf{Z} = \mathbf{X}_1 - \mathbf{X}_2, \mathbf{X}_1 - \mathbf{Y}_1, \mathbf{Y}_1 - \mathbf{Y}_2$, $\sum_{q \neq q'} Cov({Z^{(q)}}^2, {Z^{(q')}}^2)$ is of the order $\mathbf{o}(d^2)$.

(A3) There exists σ_F^2, σ_G^2 and ν^2 such that $d^{-1}\sum_{q=1}^{d} Var(X_1^{(q)}) \to \sigma_F^2$, $d^{-1}\sum_{q=1}^{d} Var(Y_1^{(q)}) \to \sigma_G^2$ and $d^{-1}\|E(\mathbf{X}_1) - E(\mathbf{Y}_1)\|^2 \to \nu^2$ as $d \to \infty$.

Hall et al. (2005) made similar assumptions to study the behavior of several classifiers in the HDLSS asymptotic regime, where the sample size is considered to be fixed and the dimension grows to infinity. Similar assumptions were also made by Biswas et al. (2014, 2015) in the context of hypothesis testing. Under assumptions $(A1) - (A3)$, following the proof of Theorem 3.1 in Biswas and Ghosh (2014), one can show that when $\nu^2 \neq 0$ or $\sigma_F^2 \neq \sigma_G^2$, powers of BG tests converge to unity as the dimension increases. Similar high dimensional consistency can be proved for Cramér tests as well. These tests based on the Euclidean distance need the two distributions to differ in their locations or average scales to perform well in the HDLSS set up. But in Examples 1 and 2, we had $\nu^2 = 0$ and $\sigma_F^2 = \sigma_G^2$.

Though each measurement variable had different distributions under F and G, pairwise Euclidean distances were unable to extract that information. That is why the tests based on T_C^{ϕ} and T_{BG}^{ϕ} had poor performance. We can capture this difference in marginals if instead of the Euclidean distance, we use other appropriate distance functions. Here we consider a class of distance functions of the form

$$\varphi_{h,\psi}(\mathbf{x},\mathbf{y}) = h\Big\{\frac{1}{d}\sum_{q=1}^{d}\psi(|x^{(q)}-y^{(q)}|^2)\Big\}, \tag{1}$$

where $h:\mathbb{R}_+\to\mathbb{R}_+$ and $\psi:\mathbb{R}_+\to\mathbb{R}_+$ are continuous, strictly increasing functions with $h(0)=\psi(0)=0$. Observe that for any $p\geq 1$, the use of $\psi(t)=t^{p/2}$ and $h(t)=t^{1/p}$ leads to the ℓ_p distance (upto a scalar multiple). We modify Cramér and BG tests using $\varphi_{h,\psi}$ in place of the Euclidean distance. So, the modified test statistics turn out to be of the form

$$\widetilde{T}_C^{h,\psi} = 2\widehat{\mu}_{h,\psi}(F,G) - \widehat{\mu}_{h,\psi}(F,F) - \widehat{\mu}_{h,\psi}(G,G) \quad \text{and}$$
$$\widetilde{T}_{BG}^{h,\psi} = \{\widehat{\mu}_{h,\psi}(F,G) - \widehat{\mu}_{h,\psi}(F,F)\}^2 + \{\widehat{\mu}_{h,\psi}(F,G) - \widehat{\mu}_{h,\psi}(G,G)\}^2,$$

where $\widehat{\mu}_{h,\psi}(F,F) = m^{-2}\sum_{1\leq i,j\leq m}\varphi_{h,\psi}(\mathbf{x}_i,\mathbf{x}_j)$, $\widehat{\mu}_{h,\psi}(G,G) = n^{-2}\sum_{1\leq i,j\leq n}\varphi_{h,\psi}(\mathbf{y}_i,\mathbf{y}_j)$ and $\widehat{\mu}_{h,\psi}(F,G) = (mn)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\varphi_{h,\psi}(\mathbf{x}_i,\mathbf{y}_j)$. We reject $\mathcal{H}_0$ for large values of the test statistic, where the cut-off is computed using the permutation principle. In this article, we consider $\psi(t)=\sqrt{t}$, $\log(1+t)$ and $1-\exp(-t)$, while $h(t)=t$ is used. The resulting test statistics will be denoted by $\widetilde{T}_C^{sqrt}, \widetilde{T}_C^{\log}, \widetilde{T}_C^{\exp}$ (for modified Cramér tests) and $\widetilde{T}_{BG}^{sqrt}, \widetilde{T}_{BG}^{\log}, \widetilde{T}_{BG}^{\exp}$ (for modified BG tests), respectively. In Figures 1 and 2 we have already seen that the modified tests based on these statistics performed well in Examples 1 and 2. Now, to carry out a theoretical investigation on their behavior in the HDLSS asymptotic regime, we make the following assumption.

(A4) For $\mathbf{Z}=\mathbf{X}_1-\mathbf{X}_2, \mathbf{Y}_1-\mathbf{Y}_2$ and $\mathbf{X}_1-\mathbf{Y}_1$, $d^{-1}\sum_{q=1}^{d}\{\psi(|Z^{(q)}|^2)-E\psi(|Z^{(q)}|^2)\}$ converges in probability to 0 as d tends to infinity.

The assumption $(A4)$ is quite common in the HDLSS literature (see, e.g., Biswas et al., 2015). The condition holds if we assume $(A1)$ and $(A2)$ with $\psi(|Z^{(q)}|^2)$ in place of ${Z^{(q)}}^2$. It holds in many other situations as well. For instance, Andrews (1988); de Jong (1995) derived some sufficient conditions based on mixingles. Under $(A4)$, for two independent random vectors $\mathbf{X}$ and $\mathbf{Y}$, $\varphi_{h,\psi}(\mathbf{X},\mathbf{Y}) - h\{d^{-1}\sum_{q=1}^{d}E\psi(|X^{(q)}-Y^{(q)}|^2)\}$ converges in probability to 0 as d diverges. Now, assume that $\mathbf{X}_1,\mathbf{X}_2\sim F$ and $\mathbf{Y}_1,\mathbf{Y}_2\sim G$ and they are independent. Define $\varphi^*_{h,\psi}(F,F) = h\{d^{-1}\sum_{q=1}^{d}E\psi(|X_1^{(q)}-X_2^{(q)}|^2)\}$, $\varphi^*_{h,\psi}(G,G) = h\{d^{-1}\sum_{q=1}^{d}E\psi(|Y_1^{(q)}-Y_2^{(q)}|^2)\}$ and $\varphi^*_{h,\psi}(F,G) = h\{d^{-1}\sum_{q=1}^{d}E\psi(|X_1^{(q)}-Y_1^{(q)}|^2)\}$. Lemma 1 shows an interesting result involving these quantities.

Lemma 1. *Suppose that h is concave and ψ has non-constant monotone derivative. Then, for any fixed $d \geq 1$, we have $e_{h,\psi}(F,G) = 2\varphi^*_{h,\psi}(F,G) - \varphi^*_{h,\psi}(F,F) - \varphi^*_{h,\psi}(G,G) \geq 0$, where the equality holds if and only if F and G have the same one-dimensional marginal distributions.*

Proof: Let $\mathbf{X}_1, \mathbf{X}_2 \sim F$ and $\mathbf{Y}_1, \mathbf{Y}_2 \sim G$ be independent random vectors. Since ψ has non-constant monotone derivative, we get $2E\psi(|X_1^{(q)} - Y_1^{(q)}|^2) - E\psi(|X_1^{(q)} - X_2^{(q)}|^2) - E\psi(|Y_1^{(q)} - Y_2^{(q)}|^2) \geq 0$ for $q = 1, \ldots, d$, where equality holds if and only if the q-th marginal distributions of F and G are identical (see Baringhaus and Franz, 2010). As a result, we have

$$\frac{2}{d}\sum_{q=1}^{d} E\psi(|X_1^{(q)} - Y_1^{(q)}|^2) - \frac{1}{d}\sum_{q=1}^{d} E\psi(|X_1^{(q)} - X_2^{(q)}|^2) - \frac{1}{d}\sum_{q=1}^{d} E\psi(|Y_1^{(q)} - Y_2^{(q)}|^2) \geq 0, \tag{2}$$

where equality holds if and only if all marginal distributions of F and G are same. Now, since h is concave and increasing, for any three real numbers a, b and c satisfying $2b - a - c \geq 0$, we get

$$h(b) \geq h\Big(\frac{a+c}{2}\Big) \geq \frac{1}{2}h(a) + \frac{1}{2}h(c) \Rightarrow 2h(b) - h(a) - h(c) \geq 0. \tag{3}$$

The proof of the lemma now follows from Equations (2) and (3). □

The quantity $e_{h,\psi}(F,G)$ can serve as a measure of separation between F and G. In fact, this can be viewed as an energy distance between F and G (see, e.g., Székely and Rizzo, 2004; Aslan and Zech, 2005). Lemma 1 shows that for every $d \geq 1$, $e_{h,\psi}(F,G)$ is positive unless the marginal distributions of F and G are identical. Therefore, it is reasonable to make the following assumption

(A5) $\liminf_{d\to\infty} e_{h,\psi}(F,G) > 0$.

The following theorem shows the high dimensional consistency (i.e., consistency in the HDLSS asymptotic regime) of the tests based on $\widetilde{T}_C^{h,\psi}$ and $\widetilde{T}_{BG}^{h,\psi}$ under this assumption.

Theorem 1. *Suppose that we have m independent observations from both F and G, which satisfy assumptions (A4) and (A5). If $\binom{2m}{m} > 2/\alpha$, then the powers of the modified tests (of level α) based on $\widetilde{T}_C^{h,\psi}$ and $\widetilde{T}_{BG}^{h,\psi}$ converge to 1 as d tends to infinity.*

Proof : For four independent random vectors $\mathbf{X}_1, \mathbf{X}_2 \sim F$ and $\mathbf{Y}_1, \mathbf{Y}_2 \sim G$, under (A4), we have $|\varphi_{h,\psi}(\mathbf{X}_1, \mathbf{X}_2) - a_d| \xrightarrow{P} 0$, $|\varphi_{h,\psi}(\mathbf{Y}_1, \mathbf{Y}_2) - c_d| \xrightarrow{P} 0$ and $|\varphi_{h,\psi}(\mathbf{X}_1, \mathbf{Y}_1) - b_d| \xrightarrow{P} 0$ as $d \to \infty$, where $a_d = \varphi^*_{h,\psi}(F,F)$, $b_d = \varphi^*_{h,\psi}(F,G)$ and $c_d = \varphi^*_{h,\psi}(G,G)$. Thus, as $d \to \infty$, $|\widetilde{T}_C^{h,\psi} - e_{h,\psi}(F,G)| \xrightarrow{P} 0$ and $|\widetilde{T}_{BG}^{h,\psi} - \gamma_{h,\psi}(F,G)| \xrightarrow{P} 0$, where $\gamma_{h,\psi}(F,G) = (b_d - a_d)^2 + (b_d - c_d)^2$.

Now let us consider the permutation distributions of $\widetilde{T}_C^{h,\psi}$ and $\widetilde{T}_{BG}^{h,\psi}$. Under a random permutation, if $(m-r)$ observations from F and r observations from G are labeled as F and the rest as G, one can

check that $|\widetilde{T}_C^{h,\psi} - \eta_r e_{h,\psi}(F,G)| \xrightarrow{P} 0$ as $d \to \infty$, where $\eta_r = 1-\{2r(m-r)/m^2+2r(m-r)/m(m-1)\} \leq 1$, and equality holds if and only if $r = 0$ or $r = m$. Therefore, when $\liminf_{d\to\infty} e_{h,\psi}(F,G) > 0$, under the permutation distribution, the probability $\Pr^*\{\widetilde{T}_C^{h,\psi} > e_{h,\psi}(F,G)\}$ converges to $2/\binom{2m}{m} < \alpha$. This proves the consistency of the modified Cramér test based on $\widetilde{T}_C^{h,\psi}$.

For the test based on $\widetilde{T}_{BG}^{h,\psi}$, under the above mentioned random permutation, we obtain $|\widetilde{T}_{BG}^{h,\psi} - \{(b_{d,r} - a_{d,r})^2 + (b_{d,r} - c_{d,r})^2\}| \xrightarrow{P} 0$ as $d \to \infty$, where $a_{d,r} = \left[\binom{m-r}{2}a_d + r(m-r)b_d + \binom{r}{2}c_d\right]/\binom{m}{2}$, $b_{d,r} = \left[r(m-r)a_d + \{r^2 + (m-r)^2\}b_d + r(m-r)c_d\right]/m^2$ and $c_{d,r} = \left[\binom{r}{2}a_d + r(m-r)b_d + \binom{m-r}{2}\right]/\binom{m}{2}$. Since $\liminf_{d\to\infty} 2b_d - a_d - c_d > 0$, following the proof of Theorem 3.1 in Biswas and Ghosh (2014), one can show that $(b_{d,r} - a_{d,r})^2 + (b_{d,r} - c_{d,r})^2 \leq (b_d - a_d)^2 + (b_d - c_d)^2 = \gamma_{h,\psi}(F,G)$, where equality holds if and only if $r = 0$ or $r = m$. So, here also, the probability under the permutation distribution $\Pr^*\{\widetilde{T}_{BG}^{h,\psi} > \gamma_{h,\psi}(F,G)\}$ converges to $2/\binom{2m}{m} < \alpha$ as $d \to \infty$, and this proves the high dimensional consistency of the modified BG test based on $\widetilde{T}_{BG}^{h,\psi}$. □

Theorem 1 shows the consistency of the modified tests when the sample sizes from the two distributions are equal. For unequal sample sizes, the calculations become pretty messy. However, as pointed out by Biswas and Ghosh (2014), the case $m \neq n$ can be seen as the case $m = n$ with some additional observations from one class. These additional observations give us more information, and as a result, the powers of the tests are expected to increase. This can be seen from Figure 3 as well, which shows the limiting p-values of the modified test based on $\widetilde{T}_C^{sqrt}$ as functions of m and n for Examples 1 and 2. From this figure one can see that the p-values decrease with increasing values of m and n. We observed the same for other modified tests as well.

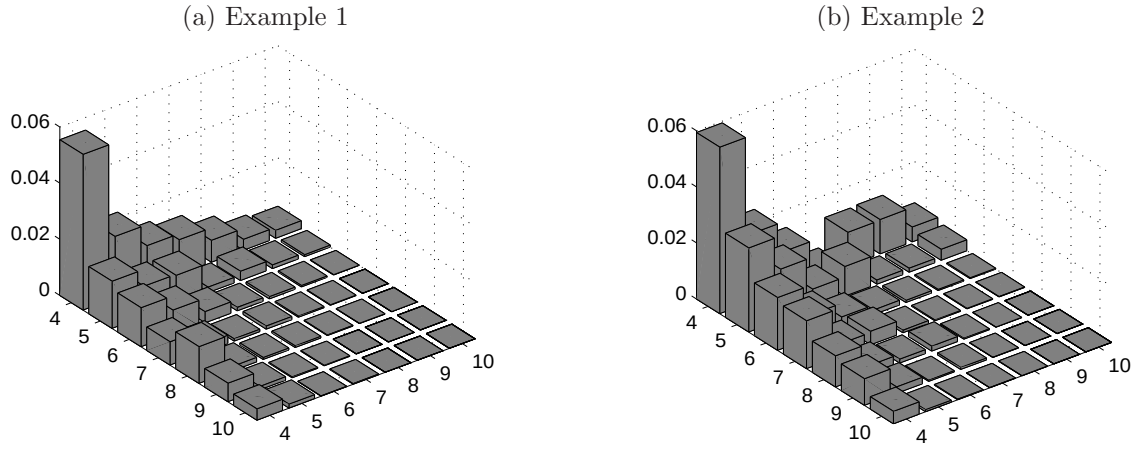

Figure 3: Limiting p-values of the modified Cramér test based on $\widetilde{T}_C^{sqrt}$ for different values of m and n.

Theorem 1 shows that even when the population distributions have the same location and the same average variance, our modified tests can distinguish between them as long as their marginal

distributions differ. This was the reason for the excellent performance of these tests in Examples 1 and 2. Now, we consider two examples, where F and G have the same one dimensional marginals.

Example 3: We consider two Gaussian distributions $\mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}_{1,d})$ and $\mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}_{2,d})$, where $\mathbf{\Sigma}_{1,d} = ((0.1^{\,|i-j|}))$ and $\mathbf{\Sigma}_{2,d} = ((0.5^{\,|i-j|}))$.

Example 4: Again we consider two Gaussian distributions $\mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}_{1,d})$ and $\mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}_{2,d})$. For $k = 1$ and 2, here $\mathbf{\Sigma}_{k,d} = ((\sigma_{ij}^k))$ is a block-diagonal matrix with $\sigma_{ii}^k = 1$ for all $i = 1, \ldots, d$, $\sigma_{2i,2i-1}^k = \sigma_{2i-1,2i}^k = \rho_k$, for all $i = 1, \ldots, \lfloor d/2 \rfloor$, and the rest are zero. We use $\rho_1 = 0.25$ and $\rho_2 = -0.25$.

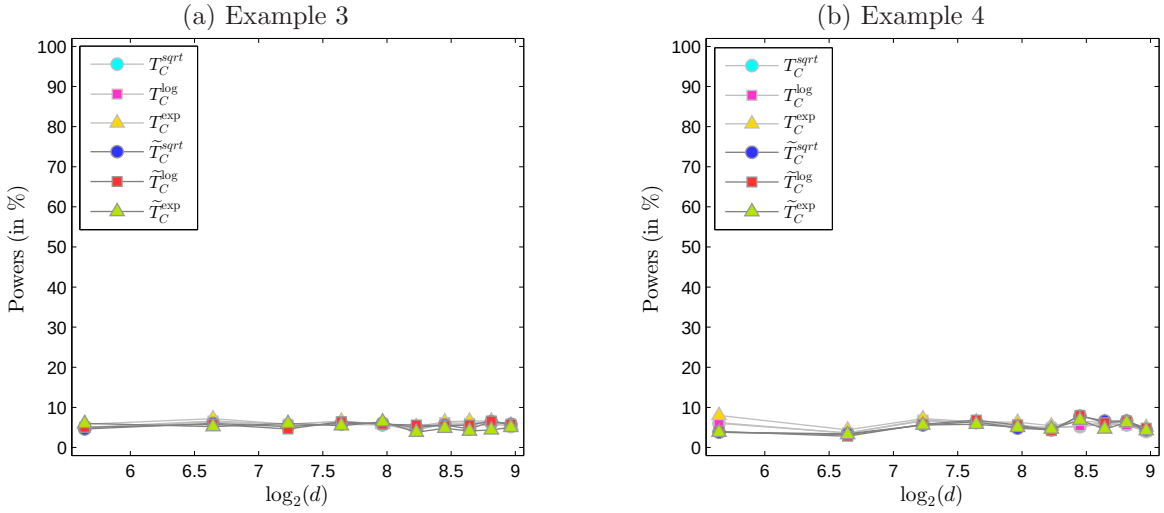

Figure 4: Powers of different Cramér tests and their modified versions based on $\varphi_{h,\psi}$.

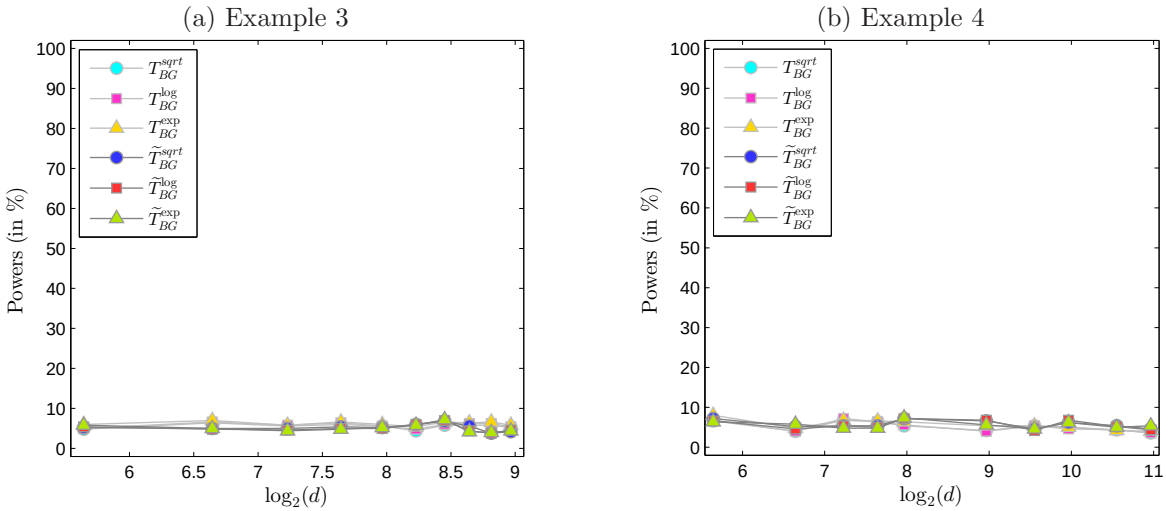

Figure 5: Powers of different BG tests and their modified versions based on $\varphi_{h,\psi}$.

In these two examples, F and G differ only in their correlation structures, and assumption $(A5)$ does not hold. Figures 4 and 5 show that all tests based on T_C^{ϕ} and T_{BG}^{ϕ} as well as their modified versions based on $\widetilde{T}_C^{h,\psi}$ and $\widetilde{T}_{BG}^{h,\psi}$ fail in these examples. This shows the necessity for further modifications of these tests. We propose such modifications in the next section.

3 Further improvement using blocking

Suppose that $\mathbf{X} \sim F$ is partitioned into D ($\leq d$) blocks as $\mathbf{X} = ({\mathbf{X}^{(1)}}^\top, \ldots, {\mathbf{X}^{(D)}}^\top)^\top$, where $\mathbf{X}^{(q)}$ follows a d_q-dimensional ($\sum_{q=1}^{D} d_q = d$) distribution F_q for $q = 1, \ldots, D$. Similarly, $\mathbf{Y} \sim G$ is partitioned as $\mathbf{Y} = ({\mathbf{Y}^{(1)}}^\top, \ldots, {\mathbf{Y}^{(D)}}^\top)^\top$, where $\mathbf{Y}^{(q)}$ follows a d_q-dimensional distribution G_q. Now, using this partition, we can define the distance between two observations $\mathbf{x}$ and $\mathbf{y}$ as

$$\varphi^B_{h,\psi}(\mathbf{x}, \mathbf{y}) = h\Big\{\frac{1}{D}\sum_{q=1}^{D} \psi\big(\|\mathbf{x}^{(q)} - \mathbf{y}^{(q)}\|^2\big)\Big\} \tag{4}$$

and construct the modified tests accordingly. The corresponding test statistics will be denoted by $\widetilde{T}^{sqrt}_{C,B}, \widetilde{T}^{\log}_{C,B}, \widetilde{T}^{\exp}_{C,B}$ and $\widetilde{T}^{sqrt}_{BG,B}, \widetilde{T}^{\log}_{BG,B}, \widetilde{T}^{\exp}_{BG,B}$, respectively. Clearly, $\varphi_{h,\psi}$ is a particular case of $\varphi^B_{h,\psi}$ with $D = d$ and $d_q = 1$ for all $q = 1, \ldots, d$. In Section 2, we have seen that the tests based on $\varphi_{h,\psi}$ failed to perform well in Examples 3 and 4, where the two distributions had the same marginals but different correlation structures. We can overcome this problem using blocks of size 2. Figures 6 and 7 show the powers of the modified tests when we used $D = d/2$, $d_q = 2$ and $\mathbf{X}^{(q)} = (X^{(2q-1)}, X^{(2q)})^\top$ for $q = 1, \ldots, D$. Clearly, these tests successfully captured the difference in the correlation structures of the two distributions to yield much improved results.

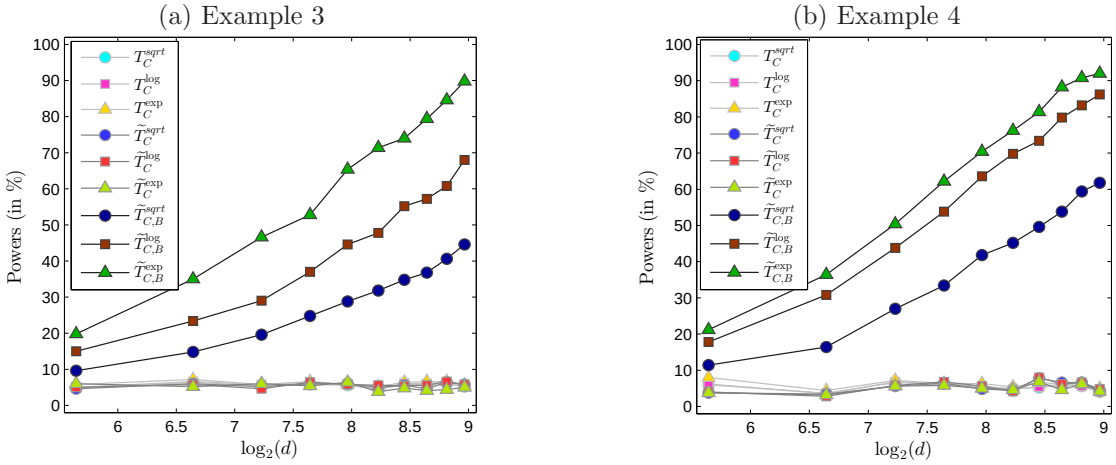

Figure 6: Powers of different Cramér tests and their modified versions based on $\varphi_{h,\psi}$ and $\varphi^B_{h,\psi}$ with blocks of size 2.

To study the high-dimensional behavior of the tests based on $\varphi^B_{h,\psi}$, we assume that the number of blocks D diverges and block sizes remain uniformly bounded as d increases. We also make the following assumption similar to $(A4)$.

(A6) Let $\mathbf{X}_1, \mathbf{X}_2 \sim F$ and $\mathbf{Y}_1, \mathbf{Y}_2 \sim G$ be independent random vectors. Then, for $\mathbf{Z} = \mathbf{X}_1 - \mathbf{X}_2$, $\mathbf{Y}_1 - \mathbf{Y}_2$ and $\mathbf{X}_1 - \mathbf{Y}_1$, $D^{-1}\sum_{q=1}^{D}\big\{\psi(\|\mathbf{Z}^{(q)}\|^2) - E\psi(\|\mathbf{Z}^{(q)}\|^2)\big\} \to 0$ in probability as $d \to \infty$.

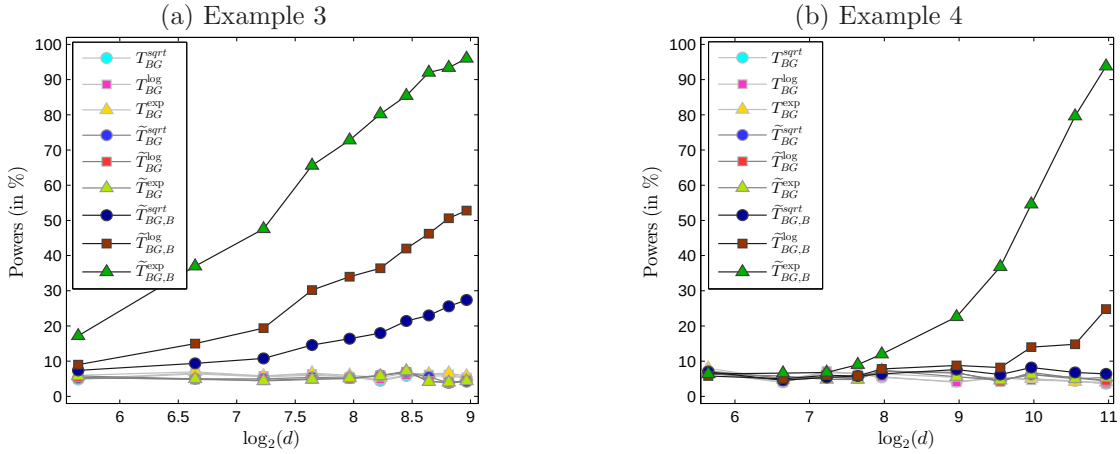

Figure 7: Powers of different BG tests and their modified versions based on $\varphi_{h,\psi}$ and $\varphi^B_{h,\psi}$ with blocks of size 2.

Under the assumption $(A6)$, for two independent random vectors $\mathbf{X}$ and $\mathbf{Y}$, we have $\varphi^B_{h,\psi}(\mathbf{X},\mathbf{Y}) - h\big\{D^{-1}\sum_{q=1}^{D} E\psi(\|\mathbf{X}^{(q)}-\mathbf{Y}^{(q)}\|^2)\big\} \to 0$ in probability as $d\to\infty$. Let $\mathbf{X}_1,\mathbf{X}_2$ be two independent copies of $\mathbf{X}\sim F$ and $\mathbf{Y}_1,\mathbf{Y}_2$ be two independent copies of $\mathbf{Y}\sim G$. As before, we define $\varphi^{B,*}_{h,\psi}(F,F) = h\big\{D^{-1}\sum_{q=1}^{D} E\psi(\|\mathbf{X}_1^{(q)}-\mathbf{X}_2^{(q)}\|^2)\big\}$, $\varphi^{B,*}_{h,\psi}(G,G) = h\big\{D^{-1}\sum_{q=1}^{D} E\psi(\|\mathbf{Y}_1^{(q)}-\mathbf{Y}_2^{(q)}\|^2)\big\}$ and $\varphi^{B,*}_{h,\psi}(F,G) = h\big\{D^{-1}\sum_{q=1}^{D} E\psi(\|\mathbf{X}_1^{(q)}-\mathbf{Y}^{(q)}\|^2)\big\}$. Results similar to Lemma 1 holds for $\varphi^{B,*}_{h,\psi}$ as well, which is stated below as Lemma 2. We skip the proof of this lemma as it is similar to the proof of Lemma 1.

Lemma 2. *Suppose that h is concave and ψ has non-constant monotone derivative. Then, for any fixed $d\geq 1$, we have $e^B_{h,\psi}(F,G) = 2\varphi^{B,*}_{h,\psi}(F,G) - \varphi^{B,*}_{h,\psi}(F,F) - \varphi^{B,*}_{h,\psi}(G,G) \geq 0$, where the equality holds if and only if $F_q = G_q$ for all $q = 1,\ldots,D$.*

Unlike $e_{h,\psi}(F,G)$, $e^B_{h,\psi}(F,G)$ can take positive values even when F and G have the same marginal distributions. The function $e^B_{h,\psi}(F,G)$ can distinguish between two populations as long as they have different block distributions. That is why the modified tests based on blocks of size 2 could capture the difference in the correlation structures of the two distributions in Examples 3 and 4. Lemma 2 shows that unless F and G have the same block distributions, for any fixed d, $e^B_{h,\psi}(F,G)$ is positive. So, it is reasonable to assume that

(A7) $\liminf_{d\to\infty} e^B_{h,\psi}(F,G) > 0$.

The following theorem shows the high dimensional consistency of the tests based on $\varphi^B_{h,\psi}$ under assumption $(A7)$ when the number of observations from the two distributions are equal. For unequal sample sizes, arguments given after the proof of Theorem 1 hold here as well. Since the proof of this theorem is similar to the proof of Theorem 1, it is omitted.

Theorem 2. *Suppose that we have m independent observations from both F and G, which satisfy assumptions $(A6)$ and $(A7)$. If $\binom{2m}{m} > 2/\alpha$, then the powers of the modified Cramér tests and BG tests (of level α) based on $\varphi^B_{h,\psi}$ converge to 1 as the dimension d diverges.*

4 Modifications of multivariate run test

Friedman and Rafsky (1979) used the idea of minimum spanning tree (MST) to construct a two-sample run test for multivariate data. They considered each of the $m+n$ observations as a vertex of an edge weighted complete graph, where the Euclidean distance between two observations was taken as the weight of the edge connecting them. They constructed the MST of this complete graph and counted the number of edges in the MST that connects observations from two different distributions. The authors suggested to reject $\mathcal{H}_0$ for smaller values of this count. This test has the large sample consistency (see, e.g., Henze and Penrose, 1999), and it can be used even for data with dimension larger than the sample size. But Biswas et al. (2014) showed that it often fails to perform well in the HDLSS set up, especially when the two distributions do not differ much in their locations. To overcome this problem, they constructed another run test using the idea of shortest Hamiltonian path (SHP). They considered the same edge weighted complete graph as above, constructed the SHP in it, and suggested to reject $\mathcal{H}_0$ if the number of runs (henceforth denoted by T_{SHP}) along that SHP is small. This test outperforms the test based on MST in a wide variety of high dimensional problems. Biswas et al. (2014) established the high dimensional consistency of this test under $(A1) - (A3)$ when $\nu^2 > 0$ or $\sigma_F^2 \neq \sigma_G^2$. However, this test based on the Euclidean distance also failed in Examples 1–4 (see Figures 8 and 9), where we had $\nu^2 = 0$ and $\sigma_F^2 = \sigma_G^2$.

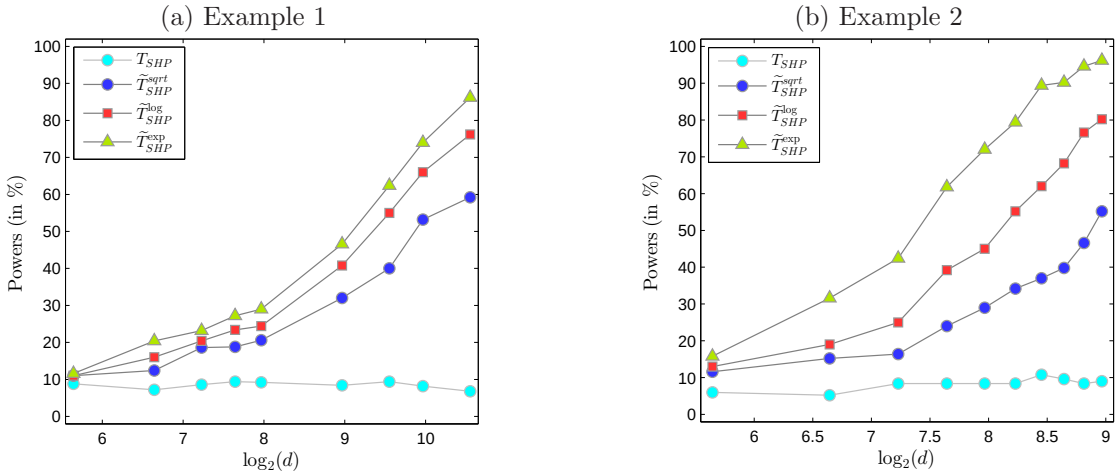

Figure 8: Powers of the SHP run test and its modified versions based on $\varphi_{h,\psi}$.

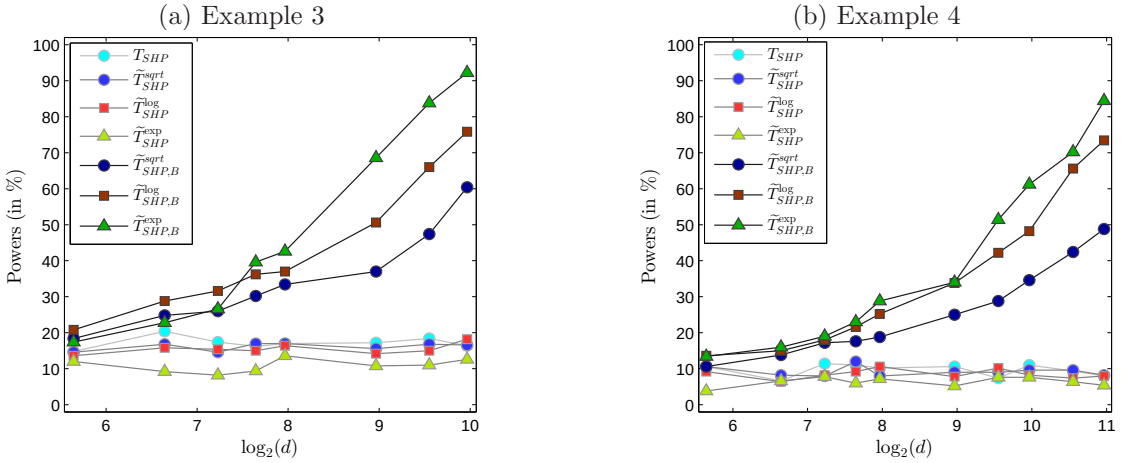

Figure 9: Powers of SHP run test and its modified versions based on $\varphi_{h,\psi}$ and $\varphi^B_{h,\psi}$ with blocks of size 2.

Figure 8 shows that instead of the Euclidean distance, when we used $\varphi_{h,\phi}$ as the edge weights of the complete graph, the resulting run tests based on SHP performed well in Examples 1 and 2 (see the curves corresponding to $\widetilde{T}^{sqrt}_{SHP}, \widetilde{T}^{\log}_{SHP}$ and $\widetilde{T}^{\exp}_{SHP}$) but they failed in Examples 3 and 4 (see Figure 9), where the two distributions had the same marginals. Using $\varphi^B_{h,\phi}$ with the same blocks of size 2 as used in Section 3, we got much improved results in these two examples (see the curves corresponding to $\widetilde{T}^{sqrt}_{SHP,B}, \widetilde{T}^{\log}_{SHP,B}$ and $\widetilde{T}^{\exp}_{SHP,B}$). This is consistent with what we observed in Sections 2 and 3, and the reasons behind different types of behavior of runs tests based on different distance functions can be explained using similar arguments as before. Results similar to Theorems 1 and 2 can also be derived for these modified SHP run tests based on $\varphi_{h,\psi}$ and $\varphi^B_{h,\psi}$, respectively.

5 Analysis of benchmark data sets

We analyzed two benchmark data sets, the `Gun-Point` data and the `Lighting-2` data, for further evaluation of different modified tests described in this article. These data sets are available at the *UCR Time Series Classification Archive* (`http://www.cs.ucr.edu/~eamonn/time_series_data/`), and they have been extensively used in the literature, mainly for supervised classification. In both of these data sets, we have reasonable separation between the two distributions. So, assuming $\mathcal{H}_0$ to be false, we compared different tests based on their empirical powers. These data sets consist of separate training and test sets. For our analysis, we merged these sets and following Biswas et al. (2014), we used random subsamples of different sizes from the whole data set keeping the proportions of observations from different distributions as close as they were in the original data set. We repeated each experiment 500 times to compute the powers of different tests, and they are shown in Figures 10 and 11.

For multivariate run tests, following Biswas et al. (2014), we used the method based on Kruskal's algorithm for finding the SHP. These run tests have the distribution-free property. In all other cases, we used conditional tests based on 1000 random permutations. For modified tests using blocks, we used blocks of size two as before, where the blocks were chosen using a data driven method based on the idea of optimal non-bipartite matching (Lu et al., 2011). Note that one would ideally like to have highly correlated variables in the same block. So, we considered a complete edge weighted graph on d vertices, where the absolute value of the correlation between a pair of variables was taken as the weight of the edge connecting them. We used the `R` package `nbpMatching` to find $\lfloor d/2 \rfloor$ disjoint pairs of variables such that the total weight between them is maximum. These pairs (plus a single variable if d is odd) were used as the D blocks.

`Gun-Point` data set comes from the video surveillance domain. It contains 100 observations from each of two distributions: 'Gun-Draw' and 'Point', where each observation consists of 150 measurements. For this data set, modified versions of the SHP run test had the best overall performance. Tests based on $\varphi_{h,\psi}$ and $\varphi^B_{h,\psi}$ had similar powers. Modified versions of Cramér tests performed better than the usual tests based on T^{ϕ}_{C}. Among these modified tests, the one based on $\widetilde{T}^{\exp}_{C,B}$ had the best performance. We observed similar phenomenon for BG tests as well, where modified tests performed better than the usual tests based on T^{ϕ}_{BG}. Among them, the test based on $\widetilde{T}^{\log}_{BG}$ and its blocked variant outperformed others.

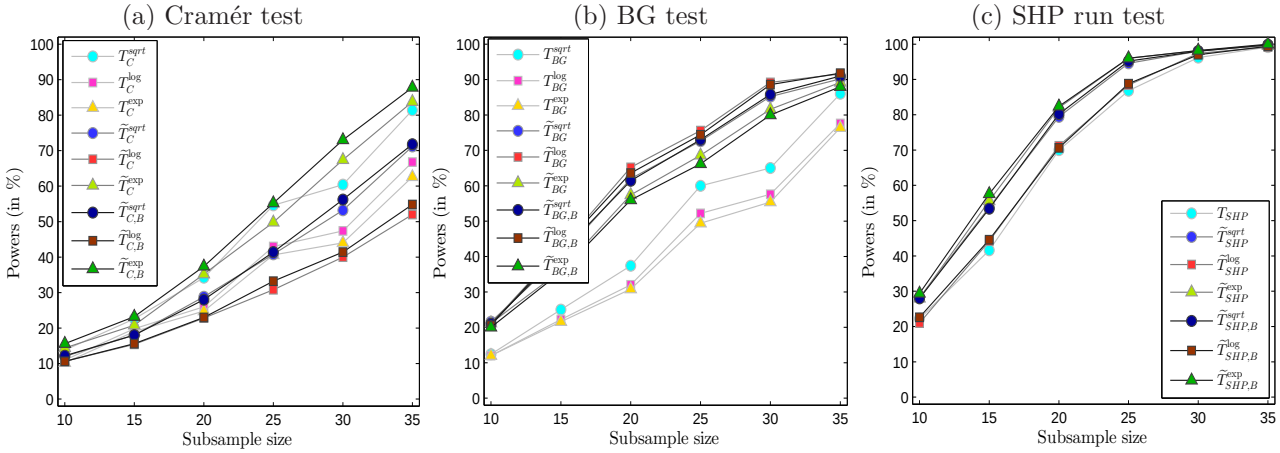

Figure 10: Powers of different tests and their modified versions for Gun-Point data.

`Lighting-2` data set contains 637-dimensional observations from two populations; 48 observations from one population and 73 from the other. Figure 11 clearly shows the superiority of the modified tests in this example. For all tests, their modified versions had much improved performance than the usual ones. Blocking led to some improvement in powers, though the difference was not that significant.

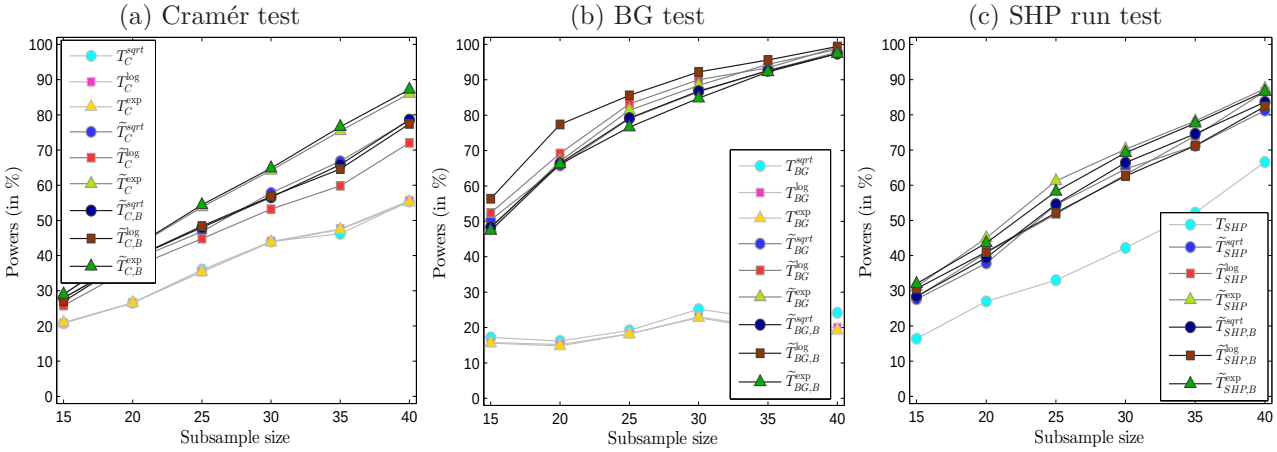

Figure 11: Powers of different tests and their modified versions for Lighting-2 data.

For the BG test, modified versions based on log transformation performed better, while for the Cramér test and the SHP run test, their modified versions based on exponential transformation had an edge.

6 Concluding remarks

Distance based methods are very useful in the context of high dimensional data analysis, and the Euclidean metric is arguably the most popular choice as the distance function. But statistical methods based on the Euclidean distance often suffer in the HDLSS set up due to the distance concentration phenomenon. In this article, we have defined a new class of distance functions and used them to modify some existing two-sample tests. These modified tests can produce superior results in a wide variety of examples, and we have amply demonstrated it using our theoretical as well as numerical results. We have also generalized the new class of distances and associated tests using blocking. It is shown that blocking can further improve the performance of the tests for a larger class of alternatives.

Throughout this article, we have used blocks of size 2, which is useful for finding differences in correlation structures. We used a data-driven method for blocking based on the idea of optimal non-bipartite matching, which worked well in all examples considered in this article. But depending on the nature of the problem, one may need to use larger block as well, and all blocks may not be of the same size. Ideally one would like to find blocks of measurement variables which are independent (or weakly dependent) of each other. Construction of a suitable data driven algorithm in this regard can be considered as an interesting topic for future research. Such blocking may further improve the performance of the resulting tests. In this article, we have modified Cramér, BG and SHP run tests using new distance functions. Similar modifications can be done for many other tests based on pairwise

distances (e.g., the adjacency test by Rosenbaum 2005, the MMD test by Gretton et al. 2012) as well. But, to save space, we omit the details in this article.

References

Anderson, T. W. (2003) *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

Andrews, D. W. K. (1988) Laws of large numbers for dependent nonidentically distributed random variables. *Econometric Theory*, **4**, 458–467.

Aslan, B. and Zech, G. (2005) New test for the multivariate two-sample problem based on the concept of minimum energy. *J. Statist. Comput. Simul.*, **75**, 109–119.

Bai, Z. and Saranadasa, H. (1996) Effect of high dimension: by an example of a two sample problem. *Statist. Sinica*, **6**, 311–329.

Baringhaus, L. and Franz, C. (2004) On a new multivariate two-sample test. *J. Multivariate Anal.*, **88**, 190–206.

— (2010) Rigid motion invariant two-sample tests. *Statist. Sinica*, **20**, 1333–1361.

Biswas, M. and Ghosh, A. K. (2014) A nonparametric two-sample test applicable to high dimensional data. *J. Multivariate Anal.*, **123**, 160–171.

Biswas, M., Mukhopadhyay, M. and Ghosh, A. K. (2014) A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika*, **101**, 913–926.

— (2015) On some exact distribution-free one-sample tests for high dimension low sample size data. *Statist. Sinica*, **25**, 1421–1435.

Chen, S. X. and Qin, Y. L. (2010) A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.*, **38**, 808–835.

Choi, K. and Marden, J. (1997) An approach to multivariate rank tests in multivariate analysis of variance. *J. Amer. Statist. Assoc.*, **92**, 1581–1590.

Ferger, D. (2000) Optimal tests for the general two sample problem. *J. Multivariate Anal.*, **74**, 1–35.

Friedman, J. H. and Rafsky, L. C. (1979) Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Statist.*, **7**, 697–717.

Ghosh, A. K. and Biswas, M. (2016) Distribution-free high-dimensional two-sample tests based on discriminating hyperplanes. *Test*, **25**, 525–547.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012) A kernel two-sample test. *J. Mach. Learn. Res.*, **13**, 723–773.

Hall, P., Marron, J. S. and Neeman, A. (2005) Geometric representation of high dimension, low sample size data. *J. Royal Statist. Soc. Ser. B*, **67**, 427–444.

Hall, P. and Tajvidi, N. (2002) Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, **89**, 359–374.

Henze, N. (1988) A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.*, **16**, 772–783.

Henze, N. and Penrose, M. D. (1999) On the multivariate runs test. *Ann. Statist.*, **27**, 290–298.

Hettmansperger, T. P., Möttönen, J. and Oja, H. (1998) Affine invariant multivariate rank tests for several samples. *Statist. Sinica*, 785–800.

Hettmansperger, T. P. and Oja, H. (1994) Affine invariant multivariate multisample sign tests. *J. Royal Statist. Soc. Ser. B*, 235–249.

de Jong, R. M. (1995) Laws of large numbers for dependent heterogeneous processes. *Econometric Theory*, **11**, 347–358.

Liu, Z. and Modarres, R. (2011) A triangle test for equality of distribution functions in high dimensions. *J. Nonparametr. Stat.*, **23**, 605–615.

Lu, B., Greevy, R., Xu, X. and Beck, C. (2011) Optimal nonbipartite matching and its statistical applications. *Amer. Statist.*, **65**, 21–30.

Maa, J. F., Pearl, D. K. and Bartoszyński, R. (1996) Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *Ann. Statist.*, **24**, 1069–1074.

Mondal, P. K., Biswas, M. and Ghosh, A. K. (2015) On high dimensional two-sample tests based on nearest neighbors. *J. Multivariate Anal.*, **141**, 168–178.

Park, J. and Ayyala, D. N. (2013) A test for the mean vector in large dimension and small samples. *J. Statist. Plann. Inf.*, **143**, 929–943.

Puri, M. and Sen, P. K. (1971) *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

Randles, R. H. and Peters, D. (1990) Multivariate rank tests for the two-sample location problem. *Comm. Statist. Theory Methods*, **19**, 4225–4238.

Rosenbaum, P. R. (2005) An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. Royal Statist. Soc. Ser. B*, **67**, 515–530.

Schilling, M. F. (1986) Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.*, **81**, 799–806.

Srivastava, M. S., Katayama, S. and Kano, Y. (2013) A two sample test in high dimensional data. *J. Multivariate Anal.*, **114**, 349–358.

Székely, G. J. and Rizzo, M. L. (2004) Testing for equal distributions in high dimension. *InterStat*, **5**.

Wei, S., Lee, C., Wichers, L., Li, G. and Marron, J. (2016) Direction-projection-permutation for high dimensional hypothesis tests. *J. Comput. Graph. Statist.*, **25**, 549–569.